1К

УДК 004.93+519.2

Б.О. Капустій[1], Б.П. Русин[2], В.А. Таянов[2]
[1]Національний університет “Львівська політехніка”, м. Львів, Україна
[2]Фізико-механічний інститут ім. Г.В. Карпенка НАН України, м. Львів, Україна
rusyn@ipm.lviv.ua, vtayanov@ipm.lviv.ua

Комбінаторна оцінка впливу зменшення інформаційного покриття класів на узагальнюючу властивість 1NN алгоритмів класифікації

У статті запропоновано комбінаторний підхід до визначення впливу зменшення розмірності класів на ймовірність правильного розпізнавання при застосуванні 1NN вирішуючого правила. Результати розпізнавання для кожного контрольного об’єкта вважаються відомими до пониження розмірів класів бази даних. Розв’язано задачу визначення ймовірності того, що правильне розпізнавання збережеться після пониження розмірності класів, а неправильне стане правильним.

Вступ

У процесі розв’язування задач розпізнавання, а також при розробці відповідних алгоритмів доцільно зменшувати складність математичних моделей алгоритмів розпізнавання з метою досягнення більшої ефективності при застосуванні цих алгоритмів на практиці. Такого можна досягти шляхом вибору підмножини найбільш інформативних ознак, застосування більш простих класифікаторів, а також зменшення розмірності класів еталонів. На сьогоднішній день проблема вибору найбільш інформативних ознак порівняно добре досліджена [1-4]. Суть селекції найбільш інформативних ознак полягає у застосуванні певного критерію, який повинен максимізувати (мінімізувати) певний показник при використанні навчаючої вибірки. Задача екстремалізації показника критерію вирішується на основі узагальненого ковзаючого контролю (cross-validation test) [5-7], який продовжує вдосконалюватися у даний час. Досліджуються також підходи, які враховують вплив кожної ознаки на внутрішню композицію алгоритму, що дає змогу більш ефективно визначати вплив сукупності ознак на узагальнюючу властивість алгоритмів розпізнавання [2]. Надалі буде розглянута задача розпізнавання з класами, що попарно не перетинаються, а процес навчання здійснюватиметься на основі прецедентної інформації [8]. У подібних випадках часто застосовують класифікатори на основі функції відстані (метричні класифікатори). В загальному ними є *k*NN класифікатори, зважені *k*NN класифікатори та класифікатори з використанням потенційних функцій [9]. Ці класифікатори більш прості і швидкісні порівняно з іншими. Сумісна процедура оптимізації метрики та найбільш інформативного набору ознак досліджена на основі різних критеріїв в [10], [11], а в [11] розроблено підхід для порівняння 1NN та *k*NN класифікаторів. В [10] на основі послідовного аналізу запропоновано підхід до визначення мінімального розміру класу, що забезпечує задані помилки класифікації 1-го та 2-го родів.

Необхідно відзначити, що дослідження, на які орієнтована дана робота, стосується здебільшого біометричних СР, хоча розроблювані підходи мають достатньо загальний характер, щоб їх застосовувати і для видів СР.

Формулювання задачі

Задача пониження інформаційного покриття класів є цільовою задачею. По-перше, при зменшенні інформаційного покриття класів підвищується швидкість процесу розпізнавання, а значить, і ефективність роботи відповідної системи. По-друге, зменшення інформаційного опису класів дає можливість зменшити вплив такого негативного явища, як перенавчання. В класичному розумінні під перенавчанням розуміється різниця оцінок результатів розпізнавання на контрольній вибірці і під час навчання. По-третє, при пониженні інформаційного покриття класів зменшується інформаційне покриття чужих класів, що виникають лише під час процесу розпізнавання і заважають його успішному здійсненню. Може трапитися ситуація, коли при пониженні інформаційного покриття результати правильного розпізнавання будуть кращими, ніж до його пониження. Це пояснюється тим, що при зменшенні інформаційного опису класів-еталонів з'являється достатньо велика ймовірність того, що образи-еталони з чужих класів, які призводили до негативних результатів розпізнавання, будуть вилучені з бази даних. Вказаний ефект також призводитиме до перенавчання алгоритму розпізнавання. Тому для побудови більш точних оцінок імовірності правильного розпізнавання необхідно проводити усереднення ймовірностей такого розпізнавання для різних розмірів класів еталонів.

При пониженні інформаційного покриття класів-еталонів можливі два варіанти. Перший передбачає правильне, а другий неправильне розпізнавання до пониження інформаційного покриття і потребують оцінки його успішності після такого пониження. Оцінка правильності розпізнавання після пониження інформаційного покриття обумовлюється співвідношенням вказаних випадків.

Формалізація та постановка задачі

Нехай X – простір об'єктів; Y – множина імен класів; $y^*: X \to Y$ – цільова функція, значення якої відомі лише на об'єктах скінченої навчаючої вибірки $X^l = (x_i, y_i)_{i=1}^l \subset X \times Y$, $y_i = y^*(x_i)$ [5]. У базі даних існують класи еталонів C_i, $i = \overline{1,n}$, причому $s_i = |C_i|$ – розміри класів. Передбачається, що розміри s_i всіх класів однакові і рівні s. Оскільки існує вибірка контрольних образів U, що подаються на розпізнавання, то загальна кількість образів, що беруть участь у процесі розпізнавання, дорівнюватиме $n * s + |U|$. Нехай оцінена частота помилок алгоритму класифікації $a = \mu(X^l)$ на навчаючій вибірці $X^l \subseteq X^L$: $\nu(a,U) = \frac{1}{|U|}\sum_{x \in U}[a(u) \neq y^*(u)]$. Задача полягає в оцінюванні величини $\widetilde{\nu}(a,U) = \frac{1}{|U|}\sum_{x \in U}[\widetilde{a}(u) \neq y^*(u)]$ при пониженні інформаційного покриття $|C_i|$ класів-еталонів, де $\widetilde{a} = \mu(X^{\widetilde{l}})$ – алгоритм, побудований на основі вибірки розміру $\widetilde{l}$. Як алгоритм класифікації використаємо найбільш простий серед метричних алгоритмів 1NN алгоритм. При такій постановці задачі найбільш придатним підходом, який можна використати для її вирішення, є комбінаторний підхід. Очевидно, що в кожному конкретному випадку пониження інформаційного покриття класів буде проводитись не обов'язково оптимальним чином, однак

загальна статистика всіх можливих понижень класів та результатів таких понижень дасть відповідь на питання про ефективність пониження інформаційного покриття класів-еталонів у цілому.

Розв'язання задачі

Представимо дані, що подаються на класифікатор a, у вигляді двійкової послідовності $\{0,1\}$, посортованої за мінімумом відстаней об'єктів бази даних від тестового об'єкта, де 1 ставляться у відповідність образам, які підтримують правильне розпізнавання (образи свого класу), а 0 – образам, які заважають такому розпізнаванню (образи чужих класів). Приклад такої послідовності поданий на рис. 1.

$$\underbrace{1111111111}_{l}\underbrace{\underbrace{000}_{m_1}\underbrace{111}_{k_1}\underbrace{00}_{m_2}\underbrace{1111}_{k_2}\underbrace{000}_{m_3}\underbrace{111}_{k_3}\ldots\underbrace{1111}_{k_n}\ldots\underbrace{000}_{m_n}}_{\{k,m\}}\ldots$$

Рисунок 1 – Модель розпізнавання при заданні початкового розміру классу у вигляді двійкової послідовності

Із наведеного рисунка видно, що послідовність образів, які підтримують розпізнавання, має розмірність $l+k$. Однак різні образи суттєво відрізняються між собою за можливостями цієї підтримки. Дійсно, при використанні 1NN правила видалення $l-1$ образів з класу-еталону не змінить результатів розпізнавання. З іншого боку, якою б довгою не була послідовність з k образів, вона не зможе підтримати розпізнавання за відсутності стратегічної послідовності розміром l і присутності послідовності розміром m.

При пониженні інформаційного покриття класу потрібно враховувати той факт, що якщо послідовність розміром l присутня у початковому класі, то у класі з меншим інформаційним покриттям s^* вона може зникнути, і навпаки, якщо її не було, то може з'явитися, однак з іншим розміром l^*.

Розглянемо 1NN правило. Визначальною перевагою даного правила є простота реалізації, а до недоліків можна віднести наступні [9]:

– *нестійкість до похибок, створених викидами у навчаючій вибірці (викидом називають об'єкт певного класу, який знаходиться в оточенні об'єктів чужих класів);*
– *повну залежність алгоритму від метрики між об'єктами та відсутність параметрів для налаштування за навчаючою вибіркою методами ковзаючого контролю або іншими;*
– *низька якість класифікації.*

Попри вказані недоліки, 1NN правило може мати суттєво кращу стійкість до ефекту пониження інформаційного покриття класів. Це пов'язано з тим, що даний алгоритм менш чутливий до розміру класів, ніж *k*NN.

Отже, можливі два випадки розпізнавання: початкове розпізнавання правильне або неправильне, і потрібно визначити ймовірність його успішності після пониження інформаційного покриття класів. Тобто для першого випадку потрібно визначити ймовірність того, що розпізнавання залишиться правильним, а для другого –

ймовірність переходу розпізнавання з категорії неправильного в категорію правильного. Подамо ймовірність правильного розпізнавання при застосуванні 1NN правила як відношення подій, які підтримують успішне розпізнавання, до загальної кількості подій:

$$P(k,l,s)=\begin{cases}\dfrac{C_{l+k}^{s^*}-C_{k}^{s^*}}{C_{l+k}^{s^*}}=1-\dfrac{k!s^*!(l+k-s^*)!}{s^*!(k-s^*)!(l+k)!}=\dfrac{k!(l+k-s^*)!}{(k-s^*)!(l+k)!}, k\geq s^*;\\ 1, \text{в іншому випадку}.\end{cases} \quad (1)$$

Пояснимо обчислення ймовірностей (1). Якщо $k<s^*$ і початкове розпізнавання було правильним, то пониження інформаційного покриття не призведе до погіршення результатів розпізнавання, тобто $P(k<s^* \mid P(s)=1)=1$. Вираз (1) означає ймовірність того, що розпізнавання буде успішним незалежно від того, яким чином буде зменшене інформаційне покриття своїх і чужих класів. Таким чином, ця ймовірність буде оцінкою зверху для точного (в сенсі комбінаторики) значення ймовірності правильного розпізнавання. Сам принцип оцінок зверху ймовірності успішного розпізнавання полягає в тому, що обчислення точного значення відповідної ймовірності вимагає застосування багатокрокового ітераційного процесу.

Уточнити значення ймовірності (1) можна шляхом введення ще однієї оцінки зверху ймовірності того, що перед послідовністю $\{k_i\}$ ($\sum_i k_i = k$) після пониження інформаційного покриття класів бази даних не буде знаходитись послідовність $\{\bigcup_j m_j\}, j<i$ ($\sum_i m_i = m$).

Після виключення з моделі (рис. 1) стратегічної послідовності вона трансформується до такого вигляду:

$$\underbrace{\underbrace{000}_{m_1}\underbrace{111}_{k_1}\underbrace{00}_{m_2}\underbrace{1111}_{k_2}\underbrace{000}_{m_3}\underbrace{111}_{k_3}\ldots\underbrace{000}_{m_n}\underbrace{1111}_{k_n}}_{\{k,m\}}\ldots$$

Рисунок 2 – Модель розпізнавання у вигляді двійкової послідовності при $\{l\}\neq\varnothing$

Таким чином, задача зводиться до визначення ймовірності успішного розпізнавання після пониження інформаційного покриття класів для випадків, коли початкове розпізнавання було неправильним. Ці ймовірності обчислюються n разів для пар послідовностей $\{m_i,k_i\}$, $i=\overline{1,n}$. Отже, на даному етапі вихідною послідовністю з усіх одиниць буде послідовність розміром k.

Означення. Показником виживання підпослідовності $\{m_i,k_i\}$ можна вважати ймовірність того, що в результаті можливих комбінацій входжень об’єктів з цієї підпослідовності в інші у ній залишиться хоча б один об’єкт з вихідної підпослідовності. Вказану ймовірність можна записати у вигляді:

$$\begin{cases}P(m_i,k_i,\{l\}=\varnothing)=\dfrac{C_{k+l-m_i}^{s^*}}{C_{k+l}^{s^*}}\left(1-\dfrac{C_{k-k_i}^{s^*}}{C_{l+k}^{s^*}}\right), k-k_i\geq s^*, m-m_i\geq s^*;\\ 1, \text{в іншому випадку}.\end{cases} \quad (2)$$

Якщо всі образи із свого класу в результаті їх сортування за величинами відстаней до тестового образу попали в межі списку $\{m,k\}$, то вираз (2) означає ймовірність того, що в цьому списку будуть знаходитись такі образи із свого і такі з чужих класів, що розпізнавання пройде успішно. Ця ймовірність обчислюється рекурсивно-ітераційним способом на основі підпослідовностей $\{k_i, m_i\}$:

$$P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing, \{m_1\} = \varnothing\right) = P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing\right) P(\{m_1\} = \varnothing) =$$

$$\frac{C_{l+k-m_1}^{s^*}}{C_{l+k}^{s^*}} \left(1 - \frac{C_{k-k_1}^{s^*}}{C_{l+k}^{s^*}}\right), k - k_1 \geq s^*, m - m_1 \geq s^*;$$

$$P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing, \{k_2\} \neq \varnothing, \{m_1\} = \varnothing, \{m_2\} = \varnothing\right) =$$

$$P\left(\{l\} \neq \varnothing, \{k_1\} \neq \varnothing, \{k_2\} \neq \varnothing\right) P(\{m_1\} = \varnothing, \{m_2\} = \varnothing) =$$

$$\frac{C_{l+k-m_1-m_2}^{s^*}}{C_{l+k}^{s^*}} \left(1 - \frac{C_{k-k_1-k_2}^{s^*}}{C_{l+k}^{s^*}}\right), k - (k_1 + k_2) \geq s^*, m - (m_1 + m_2) \geq s^*; \qquad (3)$$

………………………………………………………………………………

$$P\left(\{l\} \neq \varnothing, \{k_i\}_{i=1}^n \neq \varnothing, \{m_i\}_{i=1}^n = \varnothing\right) = P\left(\{l\} \neq \varnothing, \{k_i\}_{i=1}^n \neq \varnothing\right) P(\{m_i\}_{i=1}^n = \varnothing) =$$

$$\frac{C_{l+k-\sum_i m_i}^{s^*}}{C_{l+k}^{s^*}} \left(1 - \frac{C_{k-\sum_i k_i}^{s^*}}{C_{l+k}^{s^*}}\right), k - \sum_i k_i \geq s^*, m - \sum_i m_i \geq s^*.$$

У формулі (3) значення n визначається умовами $s - l - \sum_i k_i \geq s^*$ та $s - \sum_i m_i \geq s^*$, оскільки всі подальші ймовірності $P(\cdot)$ дорівнюватимуть 1. Добуток усіх імовірностей (3) є глобальною ймовірністю правильного розпізнавання.

Висновки

На основі запропонованого комбінаторного підходу можна всебічно дослідити найпростіший із метричних класифікаторів – 1NN класифікатор. Потенційно можливий ступінь стиску класу визначається результатами розпізнавання початкового (нестиснутого) класу. Тому за допомогою комбінаторного підходу можна оцінити вірогідність коректної роботи алгоритму розпізнавання як здатності зберігати свої параметри при зменшенні інформаційного покриття класів. Вказаний підхід не враховує ймовірності отримання тих чи інших початкових результатів розпізнавання, а працює лише на основі зареєстрованих апріорних даних.

Література

1. Ивахненко А.Г., Юрачковский Ю.П. Моделирование сложных систем по экспериментальным данным. – М.: Радио и связь, 1987. – 120 с.
2. Jiang Li, Michael T. Manry, Pramod L. Narasimha, and Changhua Yu, Feature Selection Using a Piecewise Linear Network // IEEE Transactions on Neural Network. – 2006. – Vol. 17, № 5, September. – P. 1101-1115.
3. Levner I., et al. Automated Feature Extraction for Object Recognition // In Proceedings of the Image and Vision Computing New Zealand Conference. – 2003. – P. 653-655.

4. Osborne M.R., Presnell B., and Turlach B.A. A new approach to variable selection in least squares problems // IMA Journal of Numerical Analysis. – 2000. – № 20. – P. 389-404.
5. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О.Б. Лупанова. – М.: Физматлит, 2004. – Т. 13. – С. 5-36.
6. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection // IJCAI. – 1995. – P.1137-1145.
7. M., Sukthankar R. Complete cross-validation for nearest neighbor classifiers // Proceedings of International Conference on Machine Learning. – 2000. – P. 639-646.
8. Гуров С.И. Оценка надёжности классифицирующих алгоритмов. – М.: Издательский отдел ф-та ВМиК МГУ, 2003. – 45 с.
9. Режим доступу: http:// www.ccas.ru/voron/teaching.html
10. Капустий Б.Е., Русын Б.П., Таянов В.А. Оптимизация классификаторов в условиях малых выборок // Автоматика и вычислительная техника. – 2006. – Вып. 5. – С. 25-32.
11. Капустий Б.Е., Русын Б.П., Таянов В.А. Математическая модель систем распознавания с малыми базами данных // Проблемы управления и информатики. – 2007. – № 5. – С.142-151.

Б.Е. Капустий, Б.П. Русын, В.А. Таянов
Комбинаторная оценка влияния уменьшения информационного покрытия классов на обобщающую особенность 1NN алгоритмов классификации
В работе предложен комбинаторный подход к определению влияния уменьшения размерности классов на вероятность правильного распознавания при использовании 1NN решающего правила. Результаты распознавания для каждого контрольного объекта считаются известными до понижения размеров классов базы данных. Решена задача определения вероятности того, что правильное распознавание сохранится после понижения размерности классов, а неправильное станет правильным.

B.Ye. Kapustii, B.P. Rusyn, V.A. Tayanov
In this paper the combinatorial approach for definition of the class size reduction influence on correct recognition probability when one uses 1NN classifier. The recognition results are familiar before database class size reduction for every test object. The probability that recognition system has peculiarity to retain the recognition rate after class size reduction has been determined. The probability definition task that negative recognition results after class size reduction will become positive has also been solved.

Стаття надійшла до редакції 14.01.2008.